DOCUMENT RESUME

ED 174 667 TM 009 546

AUTHOR Hayman, John; And Others
TITLE On Aggregation, Generalization, and Utility in Educational Evaluation.
PUB DATE 19 Jan 79
NOTE 25p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Collection; *Decision Making; Elementary Secondary Education; *Evaluation Methods; Evaluation Needs; *Failure Factors; Generalization; *Information Needs; Information Theory; *Information Utilization; Institutional Characteristics; Organizational Theories; *Program Evaluation
IDENTIFIERS *Cross Levels Hypothesis

ABSTRACT

The cross-levels hypothesis is presented as an explanation for program evaluation failures. It states that the usefulness of evaluation data as feedback for decision making varies inversely with the number of organizational levels between the action the data describe, and the decisions they are intended to influence. To be useful for decision making, evaluation data must meet three hierarchical information needs: syntactic, semantic, and behavioral. Syntactic need refers to the choice of organizational levels. To avoid syntactic errors, evaluators should specify their level of reference--individuals, classes, districts, states, nations--and realize that aggregating data across levels may confuse relationships among variables. Evaluation data must be on the same level as decision maker concerns, to satisfy semantic needs. Formal evaluation reports, for example, are not relevant to teachers. The behavioral need explains why decision makers are less motivated by evaluation data removed from their level--politically speaking, this information is not perceived as important to their own concerns. The cross-levels hypothesis is strongly supported by these information needs, and offers an alternative to research design or statistical procedures as an explanation for program evaluation failures. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174667

January 19, 1979

On Aggregation, Generalization, and Utility In Educational Evaluation

U S DEPARTMENT OF HEALTH EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Nick Rayder

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

by

John Hayman
Auburn University

Nick Rayder
Far West Educational Laboratory

A. Jackson Stenner
NTS Research Corporation

Doren L. Madey
NTS Research Corporation

TM009 546

The Low State of Program Evaluation

> We do not know whether or not evaluation is going to contribute more to the problems of education or more to the solutions (Stake, 1976, p. 1).

A review of the literature indicates that despite large expenditures of money and of scholarly effort, many educational program evaluation efforts to date have been failures. Orlandi and Consalve (1977, p. 3) conclude, for example, that "There is something basically wrong with program evaluation," and speak generally of its failure. Weber (1976, p. 92) notes that traditional methods of evaluation are inadequate when the evaluation problem includes a nonstructured curriculum

---

Thanks to Professor John Schmid, University of Northern Colorado and Dr. Macalia Randolf for their helpful critique of this paper.

or other highly innovative program. Rayder (1977) documents the dehumanizing aspects of many past evaluations.

A blanket statement that evaluations of educational programs have been wasted effort in most cases and downright harmful in others is certain to draw indignant rebuttals. Isolated exceptions exist, no doubt. We believe, however, that a detached, honest, and sensitive appraisal offers no conclusion but that program evaluation in education, while well-intentioned, suffers from some fundamental inadequacies.

The Cross-Levels Hypothesis

We are convinced that part of the problem relates to a very important principle which has been overlooked, for the most part, in the literature on evaluation in education. This principle concerns the practice in many evaluation studies of aggregating data across several organizational levels, and using these data to draw conclusions and make generalizations. Our hypothesis is that:

> The utility of a set of evaluation data varies inversely with the number of organizational levels between the action the data describe and the decision process they are intended to influence.

In other words, the closer a set of data is to the organizational level for which it will be used (for decision making), the more useful the data will be. For example, a program specialist in the U.S. Office of Education attempting to evaluate the results of a Title I project will gather the most useful information from descriptions of organizational behavior at the state government level.

Conversely, the principle states that the further a set of data

is from the organizational level for which it will be used (for decision making), the less useful that data will be. For example, information describing student performance in a classroom will be most useful to the classroom teacher, less useful to the principal, and still less useful to the district superintendent.

The hypothesis, considered in these terms, turns out to be very interesting, for it obviously speaks to several types of evaluation efforts, including many federally funded studies. National program evaluations tend to gather data on students at the classroom level, then apply them up through various levels until they are finally aggregated into one large data set. If the cross-levels hypothesis is valid, we can predict in advance that such an evaluation is likely to be of little value to the people it intends to inform, i.e. educational policy makers at the national level. As Bissell (1977) states in the context of Title I: "It is essential to ask whether the Title I model of a national reporting system, built upon aggregation upward of individual project data to the district, state, and federal levels, is the optimal system for addressing the diverse and often conflicting informational needs of decision makers at each of the different levels (p. 351)."

Evaluation for What Purpose

Clearly implicit in the cross-levels hypothesis is the view that evaluation is intended to influence decision making. This view of the purpose of evaluation emerged in the last decade, and currently enjoys strong support. It should be noted, however, that the "dictionary definition" of evaluation as the determination of

value is probably a good deal more widely accepted in the general education community. Many educators and program developers, at least, fear this is what evaluation means. Worthen and Sanders (1973, p. 20) note that evaluation is sometimes defined as "highly synonymous with educational measurement," and that still another view explains evaluation as the "process of comparing performance data with clearly specified objectives."

Just as evaluation can be defined in a number of different ways, the uses to which it is put vary broadly. The Stanford Evaluation Consortium (1976, p. 14) says that what is termed "evaluation" may have many purposes, including "to assess needs, to guide a 'go/no go' decision, to provide support for a decision already made, to improve plans and policies, to assist management by monitoring daily operations, (and) to test social theories." Each of these, the Consortium points out, implies "different criteria for excellence in evaluation and different, often contradictory, research tactics (p. 14)."

The cross-levels hypothesis seems most applicable to situations where evaluation results are intended to support decision making related to program improvement and/or system operation. Evaluation data as feedback, rather than as a means of establishing worth *per se*, is fundamental to the hypothesis. The hypothesis may apply in other situations, but we are not prepared to argue for such a generalization at this point.

Information and Its Purposes

Our argument for the cross-levels hypothesis begins with the

point that evaluation is a matter of producing information and transmitting it to intended users. Its purpose is to assist users in making particular types of decisions, and it must meet certain criteria to be used effectively (Stenner, 1973). The most general criterion is that information must reduce uncertainty, and if uncertainty is not reduced, no information has been exchanged (Shrode and Voich, 1976, pp. 420-421).

Weaver (1963) adds three other general criteria which information must meet to be useful as an aid to decision making. The technical or syntactic criterion deals with the proper choice of symbols to represent particular situations and with transmitting, receiving, and relating symbols accurately (Osgood, 1953, p. 699). The semantic criterion is concerned with the understanding by the receiver of the meaning of the symbols, that is, with his understanding of the relationship of the symbols to their referents (Osgood, 1953, p. 698). The behavior criterion involves the capability of a message or set of data to motivate human action.

These three aspects operate in a hierarchy; accuracy is essential to understanding, and understanding in turn is essential to effective behavior. When we say that evaluation is intended to influence decision processes, we are implying that information produced through evaluation activities meets the syntactic, the semantic, and the behavior criteria. It can fail at any of these levels. Attending only to the syntactic or semantic levels, as in application of particular methodologies in response primarily to the technical aspects of a situation, is almost certain to produce failure.

Cross-Level Evaluation and the Syntactic Criterion

Given this general background, there are a number of theoretical and pragmatic propositions which support the cross-levels hypothesis. Several relate to the syntactic criterion.

Detecting Small Signals Amidst Great Noise. -- In their review of the Handbook of Evaluation Research, the Stanford Evaluation Consortium (1976) commented on the difficulty of determining effects in social interventions:

> The relations between the processes of social intervention and evaluation has proven more complex and disconcerting than imagined. ...(The causes of social ills) are usually multiple and only a few of them are affected by any one intervention. When many factors determine outcomes, the evaluator's task is to detect a small signal amidst a great deal of noise. Moreover, when factors combine multiplicatively, a standard intervention may be necessary but insufficient, and there may be no signal to detect (pp. 12 and 13).

Educational evaluators almost invariably work in complex situations and attempt to assess the effects of various types of social interventions. The factors determining outcomes are multiple, and there is noise in the system --emanating from questionable measurement devices, for example, and from interference by other (unmeasured but relevant) variables. Under the best conditions evaluators are unlikely to produce information which can definitely pinpoint cause-and-effect relationships. The level of uncertainty about a program's operation and its effects is often as high after the evaluation has been completed and its reports delivered as it was before the effort began. See House and Associates (1978), for example, in their critique of the multimillion dollar Follow Through

evaluation.

Given the context of educational evaluation, therefore, a high probability of failure exists on the first rung of the information hierarchy ladder, i.e., in the syntactic criterion. Inaccurate symbols are likely to be produced and transmitted. As we shall see below, this situation is exacerbated when evaluative information flows across levels.

Inaccuracies in Cross-Level Generalizations. At least two problems occur when an attempt is made to collect data at one level and combine it across several levels for analysis and interpretation. The first of these involves cross-level generalizations, that is, the assumption that what the analysis seems to show at one level can be inferred upward or downward to other levels. The second problem is related directly to aggregation itself. Both errors contribute to a confusion in the syntactic function of information.

Certain concepts in General System Theory suggest that error will be increased if inappropriate units of analysis--and particularly those involving cross-level generalizations--are employed. As Miller (1975) notes:

> The universe contains a hierarchy of systems, each higher level of system being composed by systems of lower levels. Atoms are composed of particles, molecules of atoms . . . organisms or organs, groups (e.g., herds, flocks, families, teams, tribes) of organism, organizations of groups, societies of organizations, and supranational systems of societies and organizations (p. 351).

Miller states that, to avoid confusion, it is important to follow this procedural rule: "Every discussion should begin with an identification of the level of reference, and the discourse

should not change to another level without a specific statement that this is occurring (p. 21)."

In education, there are individual students, groupings in classrooms, classrooms, individual teachers, groups of teachers, schools, districts, intermediate units, states, regions and the national level. These constitute the levels, and one requirement is that we clarify our level of reference, or our choices of symbols.

We may look at units at any one of these levels -- at schools, for example, or at school districts or at state agencies. These may be described in terms of their subunits, that is, as aggregations of parts; they may be described as wholes; or they may be described in terms of the higher-level units of which they are a part, as in ascribing to individual classrooms the characteristics of the schools to which they belong. Description of an analysis unit according to its parts or according to its relation to higher-level units is liable to misinterpretation.

In a 1950 article in the American Sociological Review, Robinson (1950) discusses the issues of consistency of relations across levels of analysis. Using U.S. Census data for 1930, he shows that the "ecological correlation" between race and illiteracy was 0.95 when the units of analysis were regions of the country. When states were used as the units of analysis, the correlation dropped to 0.77, and when individuals were used, it dropped to 0.20. An "ecological fallacy" occurs when findings at the group level are generalized to the individual, that is, in inferring in this case that illiteracy and race were highly correlated for individuals.

Galtung suggests the more general term, "the fallacy of the

wrong level," to refer to this problem, and he notes:

> The fallacy can be committed working downwards, by projecting from groups or categories to individuals, or upwards, by projecting from individuals to higher units (p. 47).

We cannot conceive of a situation in which the unit of analysis relevant to a classroom teacher is the same as that relevant to a program officer in Washington, and attempts to aggregate upward from the classroom level and produce generalizations for both through the same process will almost certainly result in the wrong-level fallacy. The program officer must, by the nature of his or her task, work with regional and national generalizations, and this speaks directly to the unit of analysis which will be relevant.

Even if the relevance problem can be solved, another problem remains. Miller observes that "cross-level generalizations or hypotheses which may apply to two or more levels will, ordinarily, have greater variance than other sorts of generalizations, since they include variance among types and among individuals" (Miller, 1975, p. 352). It is not necessarily error variance which is increased, of course, but in the situations where we are attempting to detect small signals amidst great noise, introduction of new sources of variance is almost certain to compound the difficulty.

Increased Error through Aggregation. A consideration of some empirical data from Project Follow Through show concretely how aggregation affects the relationships among variables (Abt Associates, 1973). The data used in this example are the Metropolitan Achievement Test (MAT) total mathematics scores for a set of 349 second grade pupils in 43 classes in 14 Philadelphia schools for both fall 1972

and spring 1973. Of the total, 233 pupils in 32 classes in 9 schools received the Follow Through "treatment" and the remaining 116 pupils in 11 classes in five schools comprised non-Follow Through comparison groups. The relationships between fall and spring scores and between spring scores and the treatment condition were compared at different levels of aggregation, with the outcomes in the table below. The results show some sharp changes in the relationships between variables as we go from individuals to class and to school level. The regression coefficients also change with levels of aggregation.

Place Table about here

The effect of the Follow Through "treatment" on spring scores, even after controlling for initial fall scores, appears to increase as we move from individual to class to school levels of analysis. Clearly, however, the effects cannot really be changing as we move from one level to another. Nor can the differences in the relationships be due to sampling error, for the parallel analyses were based on precisely the same set of data. The only difference is that in the class and school level analyses, data were aggregated to obtain group means before the analyses were conducted. In other words, the changes apparent in the table arise solely from the effects of aggregation.

To better understand these effects, Haney (1974) conducted several other statistical analyses of these same data. He concluded that, in aggregating from individual to class level, effects

of treatment are confounded with class level effects. The same kind of confounding will occur, of course, as other aggregations occur, and to make the matter more complicated, the direction of the effect cased by confounding varies according to particular situations. That is, the confounding may cause correlations to either increase, decrease, or even change sign depending on specific circumstances (Robinson, 1950).

Now, all of this does not theoretically preclude controlling for unwanted influences. If all the relevant causal variables can be identified, accuracy can be attained. Short of doing this, however, the effects of aggregation cannot be predicted. Before we can make inferences based on data aggregated across levels of analysis, we must examine the aggregation relations and the potential manner in which they may confound relationships among variables. The criterion for success is stated by Green (1964) as follows:

> Aggregation will be said to be consistent when the use of information more detailed that that contained in the aggregates would make no difference to the results of the analysis at hand (p. 3).

In estimating the likelihood of this type of analysis being carried out successfully, we need only think of the complex situations in which most education occurs and of the vastly different local conditions which occur even within a single school district.

We can state, therefore, that the greater the diversity among the lowest level units and the greater the numbers of levels over which aggregation occurs, the greater the likelihood that analyses will be confounded by significant influences. Furthermore, since it is next to impossible to determine whether all effects have been

accounted for in an analysis, errors are likely to be frequent. The result is that the relation being studied may not be the relation the investigator thinks it is. This is the problem of "validity of relations," which Kerlinger and Pedhazur (1973, p. 81) call the "most severe" in investigations involving the relationships among variables.

Based on this, we speculate that no economically feasible scheme for evaluating a nationwide, and perhaps even statewide, program is possible, where the methodology involves aggregating individual students across various levels. In any event, the probability of serious error in meeting the syntactic criterion is quite high.

Cross-Levels Evaluation and the Semantic Criterion

What do we mean when we say that information must generate meaning on the part of the receiver? As used in semiology, the "meaning" of information refers to "the mediation process elected by (the symbols employed) and the total behavior elicited by the object(s) represented (Osgood, 1953, p. 698)." Miller (1951) explains further that "the meaning of anything to a person is the total history of his interaction with that thing." Meaning, therefore, involves both the individual's interpretative reaction to symbols received and his or her experiences with the thing symbolized.

Zais (1976, p. 230) states that meaning is related to ideas, beliefs, and attitudes, and that it "draws together the strands of (a person's) experience and makes his life comprehensible."

In an address at the 1975 Association for Supervision and Curriculum Development Convention, Arthur W. Combs observed that, "the closer an event comes to you, the greater is the degree the event has meaning for you." To say that information has meaning to a decision maker, therefore, is to imply that it relates to something in his or her own personal sphere of interest, or what Lewin (1942) refers to as the individual's "life space."

What is the sphere of interest of a principal, for example, with regard to the behavior of a teacher, and what information will meet the semantic criterion relative to the principal's concerns? Relevant variables in the behavior of Jane Doe, teacher, can be described at any number of levels. The biochemical reactions taking place in her cells could be measured and discussed. This would produce a large quantity of data. Considerable difficulty would be experienced in discerning from it the way Ms. Doe reacted to a student's request for help, however, even if the enormous syntactic problems in this example could be surmounted.

Jane Doe's biochemical processes are relevant; they are just too many levels removed from my level of reference for the semantic sense to be satisfied. I have great difficulty understanding what the data mean in terms relevant to my concern, even if all of the data have been measured and transmitted accurately. And all of it put together will not tell me all about the total behavior of the referent system, Ms. Doe.

Now, if I, as a decision maker, were concerned about the functioning of Ms. Doe's heart or some other organ, the situation would be different. The level of reference or what might be termed

the "decision unit" has changed, and what will have meaning to me as a decision maker has consequently changed. Her reaction to a student is of little concern now, except as it results in particular glands pumping and the heart palpitating. A change to still another reference level and a resulting change in meaning would occur if I were a nutritionist or a biochemist. Note that this says nothing about who produces the information; it speaks rather to what information is going to have meaning to a particular user.

To be more precise about what information will be meaningful, we noted previously that a major function of information is to reduce uncertainty. Persons in decision roles need to know periodically the status of certain "key" variables so that they can take actions which will be goal-directed.

The model, which applies at various organizational levels, is somewhat as follows:

1. The referent system has recognizable goals.
2. Characteristics which relate to the achievement of these goals are understood, at least to some extent.
3. These characteristics can be measured validly and transmitted accurately and meaningfully.
4. Different decision makers are given information on characteristics relevant to their roles and interests.
5. Based on the relevant information they make decisions and take goal-oriented actions.
6. New information on the key characteristics is then fed back, the results of the previous decisions can be assessed, and further action based on this new information can be taken. In this manner, the system is somewhat self-governing.

This model applies at all levels of self-organizing systems,

and a key point is reiteration of our previous contention that information must be relevant to the particular decision maker's concerns. Suppose, for example, that I am a classroom teacher, and I am concerned about a child's learning progress. The model then becomes very similar to what has been called "individualized" or "customized" instruction.

Information which operates in this way is likely to be highly useful to decision makers, and a major purpose of program evaluation, we believe, is to supply it. Seldom will information across more than one organization level be directly relevant to any decision making process. Decision makers must deal at their own level of abstraction and operation; those who concern themselves with details too far down or up the line will lose their effectiveness.

The Stanford Evaluation Consortium (1976) suggested that producing this type of information should be a prime concern:

> Evaluation (in their suggested model) becomes a component of the evolving program itself, rather than disinterested monitoring. . . . Formal reports to outsiders are reduced in significance, and research findings become not conclusions but updatings of the system's picture of itself (p. 18).

To summarize, information is more likely to have meaning to a decision maker if it is from an organizational level close to that at which he operates. Thus, the cross-levels hypothesis is supported by the semantic criterion, and we can conclude that errors in interpretation are likely to be introduced by aggregation and generalization across level.

Cross-Levels Evaluation and the Behavior Criterion

But what of the behavior or motivational aspect of information? Recall that this refers to the capability of a message or set of data to motivate action on the part of the receiver. To consider this apsect, we need to consider again the many ways in which evaluative information may be used.

The Stanford Evaluation Consortium noted that there is almost always a political situation with which educational evaluation is concerned. Evaluators like to think of themselves as producing data which will lead to rational decision making, and they are greatly disappointed when their output seems to be ignored.

They are often tempted to conclude that decisions about education are mostly irrational or, at best, non-rational. This is a serious mistake, for all decisions are rational from the point of view of the person making them. It is a matter, rather, of whose priorities are operating. "Certainly, the evaluator who believes that the introduction of evaluation facts and findings will make an argument less political is dangerously deluded (Stanford, 1976, p. 12)."

What information will influence a decision maker? Obviously, it will be information which the decision maker perceives as important to his own concerns. Guba and Clarke (1975, pp. 56-61) noted that practicing educators will attend only to those items which they perceive as directly relevant to their operational problems. They further observed that one of the major influences on an individual's behavior in an organization is his own needs, no matter what role he is playing.

The problem is something like determining in behavioral analysis what will serve as a reinforcer. There is no doubt that human behavior is conditioned; what is a reinforcer for me may not be one for you, however, and different reinforcers have different operational meanings for different persons. A particular decision maker will be motivated by what he perceives or has been conditioned to accept as in his own best interests, as related directly to the goals he is seeking to achieve and/or as directly relevant to his own experiences.

The cross-levels concept supports this idea for, in most cases, human beings will be less motivated by information removed from the level at which they are operating. To illustrate, consider our concerns for the different generations of our families. We take a great deal of interest in ourselves. We are also interested in our children and our parents. Our interest lags a good deal when we shift to grandparents and grandchildren. Heilbroner (1975) refers to this phenomenon as the absence of a "collective bond of identity" with past and future generations (p. 115).

For another example, consider program officers and section chiefs in different federal agencies in Washington. What are their real concerns, and what kind of information will motivate them? Most of the ones we have met really do care about education and its improvement. What interests them much more directly than the condition of a kindergarten child in Alabama or California or Montana, however, is the treatment of their program in the budgeting process and in the power games they play. What do Congressmen care about and react to? Is it any wonder that the mountains of

Title I or Follow Through evaluation documents are so much pabulum to them? This is not to take pot shots at federal types, for we all are motivated in the same way. It is, rather, to point out the cogency of the cross-levels concept with regard to the behavior criterion of information.

Suggestions for Action

To restate the hypothesis:

> The utility of a set of evaluation data varies inversely with the number of organizational levels between the action the data describe and the decision process they are intended to influence.

The validity of this hypothesis has been argued from different perspectives, and we feel it is strongly supported. Certainly the experiences of the last decade in program evaluation beg for explanation other than that commonly given--that the objectives were not clearly stated, that the measuring instruments were inadequate, that incorrect statistical procedures were used, that the evaluator got into the act after the fact so that little design was possible, etc.

We have shown that in aggregation of data across several organizational levels, the syntactic, the semantic, and the behavior criteria are each threatened. Note that, for an evaluation to succeed, all of these criteria must be satisfied. If we allow the generous estimate that the probability of each being met in an aggregation design is .50, then the probability of such an evaluation succeeding is .50 x .50 x .50 = .125, hardly a figure to inspire

confidence.

The hypothesis certainly is plausible, and among its other features, it makes common sense. It could very well explain where part of the problem in program evaluation lies, and for this reason, it deserves serious consideration.

Note that the hypothesis says nothing about statistical procedures, nor does it in any way address the need for rigor and for the strongest designs that are possible. It refers only to the need for proper units of analysis, such as the child, the classroom, the school, age groups, etc. In its definition of which units are proper, it suggests major changes in certain types of evaluation activity.

To the extent that the hypothesis is supportable, evaluators in planning their activities should consider first the organizational levels of the decision maker(s) their results are intended to serve. For each decision maker, analysis units should be chosen at his organizational level or at immediately-adjacent levels. The evaluation should deal with information on the behavior of units at these levels. For example,

- If the intent is to help decision makers within the U.S. Office of Education, the evaluation should use actions at the state agency level as its unit of analysis.
- If the intent is to serve program developers and/or directors at the school system level, then the school would be the proper unit of analysis.

This is not to say that a single evaluator cannot undertake evaluation of a nationwide program. It says, rather, that the analysis units which are appropriate depend directly on the particular

decision makers to be served. The purpose of an evaluation might be to provide local planners a basis for determining which of a number of program approaches would be likely to succeed best in their particular circumstances. A national contractor could very well perform this task, but notice that it would not be through aggregation. The case study approach described by Stake (1976, p. 96) and the explanatory observation study approach described by Cooley (1978) illustrate promising alternatives in this situation.

The Cross-Levels Hypothesis challenges some of the concepts and practices of what the Stanford Consortium refers to as the "mainstream of evaluation" (1976, pp. 17-19). If the cross-levels hypothesis finds support, the implications are substantial for some of the currently showcased practices of methodology.

Table

Relationships Between Fall Test (F), Spring Test (S) and Treatment (T) at Three Levels of Aggregation

	Level		
	Individual	Class	School
Correlations			
r_{FS}	0.529	0.423	0.356
r_{FT}	-0.016	0.052	0.148
r_{ST}	0.070	0.213	0.379
Regressions			
b_{SF}	0.90	0.63	0.59
$b_{ST.F}$	3.25	4.13	4.52
Number of Units	349	43	14

References

Abt Associates. National Evaluation of Project Follow Through, 1971-1972. (Untitled.) Cambridge, Massachusetts: Abt Associates, 1973. (2 volumes.)

Bissell, Joan S. "Review of Evaluation and Reform: The Elementary and Secondary Act of 1965." American Educational Research Journal. 14:3 Summer 1977. pp. 350-354.

Brown, Warren B. "Systems, Boundaries, and Information Flow." In Organizational Systems: General Systems Approaches in Complex Organizations. (Frank Baker, ed.) Homewood, Illinois: Richard D. Irwin, Inc., 1973.

Combs, Arthur W. "Humanistic Imperatives in an Accountable School System." Paper delivered at the Annual Convention of the Association for Supervision and Curriculum Development. New Orleans, 1975.

Cooley, William W. "Explanatory Observational Studies." AERA Invited Address, Toronto, Canada, 1978.

Galtung, J. Theory and Methods of Social Research. New York: Columbia University Press.

Green, H.A.J. Aggregation in Economic Analysis: An Introductory Survey. Princeton, N.J.: Princeton University Press, 1964.

Guba, Egon, and David Clark. "The Configurational Perspective: "A View of Educational Knowledge Production and Utilization." Washington, DC: Council for Educational Research and Development, 1975.

Haney, W. "Units of Analysis Issues in The Evaluation of Project Follow Through." Unpublished report, Huron Institute, Cambridge, Massachusetts, 1974.

Heilbroner, Robert L. An Inquiry into the Human Prospect. New York: W.W. Norton and Company, Inc., 1975.

House, Ernest R., Glass, Gene V., McLean, Leslie D. and Walker, Decker F. "Critiquing a Follow Through Evaluation." Phi Delta Kappan, March, 1978, pp. 473-474.

Kerlinger, Fred N., and Elazar J. Pedhazur. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, Incorporated. 1973.

Lewin, Kurt. "Field Theory of Learning." In The Psychology of Learning. Chicago: University of Chicago Press, 1942.

McLaughlin, Milbrey W. Evaluation and Reform: The Elementary and Secondary Education Act of 1965, Title I. Cambridge, Massachusetts: Ballinger Publishing Company, 1975.

Miller, George A. Language and Communication. New York: McGraw-Hill Book Company, Inc., 1951.

Orlandi, Lisanio R., and Robert W. Consalve. "Blaming the Victim of Evaluation." The CEDR Quarterly. 10:2, Summer 1977, pp. 3, 7.

Osgood, Charles E. Theory and Method in Experimental Psychology. New York: Oxford University Press, 1953.

Rayder, Nick, ed. Educational Evaluation and Public Policy. University of North Dakota Press, Grand Forks, North Dakota, 1977.

Robinson, W.S. "Ecological Correlations and the Behavior of Individuals." American Sociological Review. Vol. 15, 1950, pp. 351-357.

Shrode, William A. and Dan Voich, Jr. Organization and Management: Basic Systems Concepts. Homewood, Illinois: Richard D. Irwin, Incorporated., 1974.

Stake, Robert. "Evaluation Design, Instrumentation, Data Collection and Analysis of Data." Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1976. (mimeo.)

Stake, Robert. "The Case Study Method in Social Inquiry." Educational Researcher. 7:2, February 1978, pp. 5-8.

Stanford Evaluation Consortium. "Handbook of Evaluation Research." Educational Researcher. 5:10, November 1976, pp. 9-19.

Stenner, A. Jackson. Information Based Evaluation: A Design Procedure. Arlington, Virginia: IBEX Corporation, 1973.

Weaver, Warren. "Recent Contributions to the Mathematical Theory of Communications." In The Mathematical Theory of communication. (Claude E. Shannon and Warren Weaver, eds.) Urbana, Illinois: University of Illinois Press, 1963.

Wever, Margaret B. "Alternative Education: An Evaluation Dilemma." The Educational Forum. XLI:1, November 1976.

Worthen, Blaine R. and James R. Sanders. Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones

Publishing Company, 1973.

Zais, Robert S. Curriculum Principles and Foundations. New York: Thomas Y. Crowell Company, 1976.